SOK-1005: Project Assignment Spring 2025

The project assignment can be carried out alone, or in a group of up to two people. The goal of the project is for you to take us through the entire data science process of reading, processing and presenting your analysis of the data in this assignment. The files that make up the delivery of the project will be in a GitHub repository. The link to this repository must be submitted on Wiseflow by the deadline. If it is a group submission, the members of the group must submit together. Supplementary guidance will be forthcoming. The project must be carried out in R or Python, or both languages. An html file must be generated from a Quarto file, which is to be considered as the actual answer.

The data for this report should be from a data file described in Task 1. This data file will also be delivered as part of the project, so that the html file is reproducible from Quarto and the data file. If the project consists of several files, a readme file must be submitted that explains how the answer is structured. Quality and discussion count over quantity. If you create a figure or table, it is the reason and description of this that is important. There is no need to repeat this figure/table several times. There are no restrictions on length but imagine that the answer should be read by people who would assess whether your company was awarded the assignment described in the text.

Background

In the task below, you are given the opportunity to analyze sales data across various time dimensions—weekday, monthly, and yearly—across multiple stores. The primary objective is to examine how sales patterns evolve over time and to explore how store-specific demographic characteristics, such as average age and income, relate to changes in sales performance. Each task contributes equally to your final grade (1/4 each), but you are not restricted to the tasks outlined. Think of yourself as a data analyst within the company, tasked with developing a comprehensive reporting system for sales across all stores. Use the data to demonstrate the full extent of your analytical capabilities and insights.

Note: Use R/Python code that merges the datasets into one large dataset. The data is described in a separate document, "Data for Project Task Spring 2025".

Task 1:

a) Using the daily data, compute the total (aggregate) sales across all stores for each weekday (Monday to Sunday) in 1990 and 1996. Create a visualization, such as a line or bar chart, to compare weekday sales patterns between the two years. Based on the plot, describe any noticeable trends or changes over time. Are certain weekdays more prominent in 1996 than in 1990? Comment briefly on any shifts in the relative importance of specific weekdays.

b) Calculate the sales percentage change for each store for each weekday (Monday to Sunday) between 1990 and 1996. Then, create separate histograms/density plot for each weekday to visualize the distribution of these changes across stores (NB: Consider sales change only between -100% and 100% while visualizing the histogram/density plot for simplicity). Analyze the variability in sales for each weekday. Do you observe any patterns or significant differences in sales performance across weekdays?

c) Do store-level demographic factors (such as age, income, household size, etc.) explain differences in sales percentage changes across stores on weekdays? Use visualization, regression analysis, or correlation analysis or a combination of them to examine the relationship between sales changes and demographic variables. Summarize your findings clearly.

   For instance, one could perform a regression analysis where the weekday sales percentage change serves as the dependent variable, and demographic variables such as age, income, or household size are used as explanatory (independent) variables. That is:

```
fit < - lm(weekday sales percentage change = age + income +….., data= df)
summary(fit)
```

Task 2

Aggregate the daily data to monthly data, and answer the following questions:

a) Calculate the average monthly aggregate sales across all stores for the period from 1990 to 1996. Then, determine which month has the highest and lowest average aggregate sales.

b) Are monthly aggregate sales patterns consistent over the course of the year 1990 to 1996? In other words, visualize the overall seasonal trend by plotting the average monthly sales over the course of the year, and describe any noticeable seasonal patterns, such as peaks or dips in sales during specific months.

c) Pick the data in 1990 and 1996 only and calculate the month-to-month percentage change in sales for each store (i.e., comparing January 1990 to January 1996, February 1990 to February 1996, etc.).

Plot a single histogram for each month showing the percentage sales changes across all stores. Analyze the histograms to identify months where most stores experienced positive growth, where there was greater variation in sales growth, and whether there were months with predominantly negative growth.

d) Do some stores experience larger or smaller month-to-month sales changes between 1990 and 1996? Why? Are any of the demographic factors (such as age, income, household size, etc.) significantly related to the sales growth fluctuations across stores?
Use visualization, correlation analysis, regression analysis, or a combination of these techniques to answer the question.
For instance, one could perform a regression analysis where the month-to-month percentage sales changes serve as the dependent variable, and demographic variables such as age, income, etc. are used as explanatory variables. That is:

```
fit < - lm(month-to-month percentage sales = age + income +….., data= df)
summary(fit)
```

Task 3

Aggregate the data on a yearly basis to capture long-term trends and answer the following questions:

a) Calculate the average yearly aggregate sales across all stores for the period from 1990 to 1996. Then, determine which year has the highest and lowest average aggregate sales.
b) Calculate the percentage change in total sales for each store between the years 1990 and 1996. How is the percentage change in total sales distributed across all stores? Did most stores experience similar growth, or are there large differences?
c) Which stores experienced the highest (lowest) percentage change of sales between 1990 and 1996? How do the stores compare to each other in terms of growth?
d) Did some stores experience greater sales growth than others? Why? Investigate whether and how store-level demographic variables (such as age, income, household size, etc.) help explain differences in sales growth across stores.
Apply a regression, correlation or another suitable method (such as visualization). Clearly identify the key predictors of sales growth and briefly discuss the possible business implications of your findings.

Task 4.

Can the data be used to inform future sales planning? If group management is considering opening a new store, how can the data help identify the most promising location?

Drawing on the analysis above and/or conducting additional analysis provide a thorough and well-reasoned response to this question, focusing on how insights from sales trends and demographic patterns can guide strategic decision-making.

Project Presentation

A 10-minute presentation of the project will be carried out on Thursday 22 May. In this presentation, you/you will focus on the "innovative" analytical measures that led the group to choose IA as supplier of this service. An approved project presentation is required for the project assignment to be assessed.

Evaluation

The purpose of the project assignment is to give you training in:

1. ...write clearly
2. ....use data to create figures and tables
3. ...deliver according to a description

The submission is mainly judged on the content of the report, but it is also expected that the R/Python code is well documented, neat and clear. Remember that the delivery must be readable both by the general manager of a company, as well as by the corporate management. A good answer answers the assignment specifically, provides clear definitions of relevant data and economic concepts, discusses relationships based on economic theory, and explains the supplementary content of figures and tables. Remember that self-explanatory titles and categories of characters are important.

Submission Deadline

The deadline for submission of the project assignment is Monday 05 June at 16:00. Remember that the files that make up the delivery of the project should be in a GitHub repository. The link to this repository must be submitted on Wiseflow by the deadline. If it is a group submission, the members of the group must submit together. Supplementary guidance will be forthcoming.

Important information:

1. This is an individual assignment or may be completed in pairs, but no collaboration is allowed between different students or groups.
2. Be aware that copying other works is not allowed and will be considered cheating.

3. You are allowed to use any kind of sources (R codes from lectures, online resources) and offcourse AI tools. However, if ChatGPT or other AI tools are used in your work, you must describe it, in the appendix of your report, see **this link** for further information.